图书情报工作
第 64 卷 第 8 期 2020 年 4 月

基于谱聚类的虚拟健康社区知识聚合方法研究

■ 张海涛[1,2] 宋拓[1] 周红磊[1] 张鑫蕊[1]

[1] 吉林大学管理学院 长春 130022 [2] 吉林大学信息资源研究中心 长春 130022

摘 要：[目的/意义]改善虚拟健康社区知识聚合质量,为虚拟健康社区服务提供技术方法支持。[方法/过程]运用谱聚类方法对虚拟健康社区中的知识进行抽取,利用概念相似度计算得到知识主题相似度矩阵,根据该相似度矩阵进行谱聚类。[结果/结论]利用好大夫在线健康咨询平台发布的信息作为数据来源进行方法验证。结果表明,当聚类个数为 5 时,本文提出的方法得分值最高。通过谱聚类的方法充分挖掘虚拟健康社区潜在信息,改善了知识聚合质量,为知识聚合和知识服务提供了一条新途径。

关键词：虚拟健康社区 知识聚合 谱聚类 相似度

分类号：G250

DOI：10.13266/j.issn.0252-3116.2020.08.015

1 引言

党的十九大报告指出,我国社会主要矛盾已经转化为人民日益增长的美好生活需要和不平衡不充分的发展之间的矛盾。人们对健康生活的向往是健康医疗产业发展的根本动力[1]。随着居民可支配收入的不断增长,人们对疾病诊治提出了更高的要求,也对保持健康提出了新的期待。远程医疗、健康管理、科学养老等领域推动了健康医疗产业的不断革新,同时衍生出互联网医疗等许多新生事物,受到社会各界的广泛关注。随着人们对健康问题的日益关注以及基于用户生成内容的在线社区的出现,越来越多的用户利用虚拟健康社区交流健康信息和意见。虚拟健康社区旨在通过降低医疗成本、充分利用现有资源和为患者提供更多样化的沟通交流渠道来提供更好的治疗[2]。虚拟健康社区,已经成为用户交流健康经验的主流平台。

然而当前对于虚拟健康社区知识聚合及服务方面的研究较少,如何揭示和挖掘虚拟健康社区知识帖子中蕴含的知识,实现面向用户需求的知识聚合,创新虚拟健康社区的知识服务模式,提高知识服务能力和质量,成为困扰虚拟健康社区开展知识服务的首要问题。本文针对虚拟健康社区特点,结合国内外相关研究成果,采用谱聚类的方法对虚拟健康社区内容知识进行聚合,并通过实证样本数据对虚拟健康社区知识聚合方法进行验证。虚拟健康社区发展迅速且潜力巨大,本文的研究为虚拟健康社区知识聚合提供新的研究视角,对于发现虚拟健康社区的问题和不足、提升虚拟健康服务的用户满意度、促进虚拟健康社区的可持续发展具有重要意义。

2 相关概念及研究

2.1 知识聚合

“聚合”概念源于化学术语,本意是将分散的单体小分子结构通过链接关系聚集成大分子结构的过程[3]。图书情报领域学者也对聚合这一概念进行了深入的研究,聚合的研究对象从数据逐渐过渡到知识领域。李亚婷从知识服务过程的角度对知识聚合进行了定义,在知识服务的过程中,将无序的、分散的知识进行凝聚,可以发现知识单元间的关联、形成有机的知识体系[4]。王敬东认为知识聚合是一个知识聚类分析的过程,对知识聚合后,知识内涵更加丰富,使得决策过程更有意义[5]。贯君、毕强等认为知识聚合是为了构建多维又相互关联的知识体系,可以通过数据挖掘、人工智能等方法提取知识单元以及知识单元之间的内在关系[6]。李洁认为知识聚合实现的过程包括知识的聚集与知识的统合,可以通过关联和聚类对海量的信息资源进行筛选和挖掘,从而得到知识的智能融合[7]。

通过上述定义可以看出,知识聚合是运用数据挖

作者简介：张海涛(ORCID:0000-0002-9421-8187),教授,博士生导师,E-mail:zhtinfo@126.com;宋拓(ORCID:0000-0003-1282-1861),博士研究生;周红磊(ORCID0000-0002-9732-8138),硕士研究生;张鑫蕊(ORCID:0000-0001-9413-6109),硕士研究生。

收稿日期:2019-09-01 修回日期:2019-11-12 本文起止页码:134-140 本文责任编辑:易飞

掘、语义技术等人工智能手段和方法,通过分析知识的特征,将无序的、分散的知识进行重新组织和筛选,进一步发现知识之间的关联,并形成有机的知识体系,从而为用户提供具有针对性、完整性、系统性的服务,使得知识可以被高效利用的过程。针对不同的知识形态,可以使用不同的聚合方法。目前主流的知识聚合方法包括基于语义增强的知识聚合方法、多维知识聚合方法和基于类聚的知识聚合方法[8]。基于语义增强的聚合方法可以解决知识聚合过程中语义缺失的问题,一般与概念进行关联,或者使用语义标签。多维知识聚合方法是利用“用户-资源-标签”的多维划分方法进行知识的导航和推荐。Folksonomy 和社会网络分析是常用的多维知识聚合方法。基于类聚的知识聚合方法是按照知识关联的程度将知识进行关联和聚合,例如文本聚类、标签聚类都是常用的聚合方法。

2.2 虚拟健康社区知识聚合

网络社区知识聚合有其发展起源,按照时间先后分别呈现出聚合层次逐渐深入(从信息聚合到知识聚合)和聚合场景从特殊到一般(从馆藏资源知识聚合延伸到学术型社区知识聚合,再扩展到一般性的网络社区知识聚合)的逻辑顺序。研究层次的深入和研究场景的延展一方面使得面向网络社区开展知识聚合有其必然性,一方面又为其研究提供了坚实的基础[8]。

张连峰等结合学术社区用户的相关知识需求分析,建立了融合主题与 SECI 模型的虚拟学术社区知识聚合整体模型构架[9]。胡媛等基于社区中用户交流行为与用户需求设计了基于知识聚合的数字图书馆社区服务推送系统[10]。商宪丽等基于标签共现的方法设计了学术博客知识资源聚合的方法[11]。K. Liang 等分析了碎片化学习行为的特点,根据学习者的个体学习需求重新对在线教育中的知识进行聚合,从而指导学习者充分利用分散的时间来获得准确、有意义的知识内容[12]。V. Tarko 等介绍了基于流程的知识聚合和集成方法,并基于聚合机制设计了依赖于元专家和计算机算法的聚合系统,以此为基础,为知识聚合提供了工具,并探讨了构建“虚拟智库”的可能性[13]。M. Ritou 等提出了一种基于知识的多层次聚合策略来支持决策,通过对知识进行聚合的方法智能生成有意义的数据,并利用航空业的制造流程中产生的数据验证了策略的有用性,从而对制造过程进行辅助决策[14]。J. Oostermana 等研究了不同的知识提取和聚合配置如何影响艺术品注释的识别,利用众包的方法对艺术品局部注释进行自动聚合,从而方便艺术品的访问和检索[15]。虚拟健康社区中包含大量的知识单元,在各个知识单元之间存在潜在的联系和影响,揭示和发现用户生成答案的关联知识是实现答案知识的有效组织、管理和知识发现的基础。

2.3 谱聚类

聚类是一种无监督学习方法,是一种发现和探索事物内在联系的有效手段,并被广泛应用到各个领域中。聚类不需要先验知识,通过聚类分析可以将具有相似性的对象划分成类簇,使得簇内对象尽可能相似,簇外对象尽可能不同。通过聚类可以将不同的知识进行区分,将知识聚类划分为类簇后,用户可以通过聚类结果提取出知识。K-means 算法、FCM 算法、PAM 算法、PF 算法、SM 算法和 NJW 算法等聚类方法都是经典的聚类方法,可以有效对球状簇进行划分,但是对于非凸形状的簇并不适合,且容易陷入局部最优解。谱聚类作为一种基于图论的聚类方法,可以有效发现任意形状的簇结构,并且收敛于全局最优解。谱聚类算法将各个数据作为图的顶点,将相似度作为连接各个顶点的权重,计算得到顶点的邻接矩阵和相邻矩阵,将其转化为拉普拉斯矩阵之后,求得特征值以及其对应的特征向量,从而达到对数据降维以及划分的目的。R. Janani 等将谱聚类与群体优化结合,用来处理海量文本文件,通过标准数据集进行验证,并且与球面 k 均值、期望最大化法和标准粒子群算法进行比较,发现该算法比其他聚类算法具有更好的聚类精度[16]。X. Li 利用特征值差和正交特征向量对谱聚类进行改进,实现了聚类数的自动确定,利用该算法对二维评分矩阵中用户和项目进行聚类,对聚类后的评分矩阵进行分解,得到共享评分矩阵,仿真结果表明,与其他 8 种传统的协同过滤方法相比,该方法能有效地提高推荐精度和泛化能力[17]。

虚拟健康社区发布的帖子中包含大量的医学知识,并且各个帖子中包含内部联系。目前的虚拟社区聚合方法采用传统的聚类方法使得聚合后的知识缺乏语义;或者因需要构建本体而耗费大量的精力。而谱聚类的方法可以利用帖子中知识的相似度进行聚类,从而增强了知识的语义关联。通过对虚拟健康社区的帖子聚类可以有效建立帖子之间的关系,对先验知识进行聚类分析,发现各个文档中包含的知识,使得聚合后的结果更加丰富。将虚拟健康社区中的知识进行聚合就是为了满足用户知识需求,采用相关计算机方法对答案中处于离散分布状态的知识单元以及其之间的关系进行挖掘和提取,实现社区的关联知识单元的紧

密联系和有序化组织。通过这种方式,可以为虚拟健康社区用户提供满足用户个性化需求的知识推荐和知识发现服务,进一步提升虚拟健康社区用户的服务质量和用户体验。

3 计算方法及过程

获取知识是知识聚合的前提,概念是知识的核心单元。在进行虚拟社区知识聚合的过程中需要对文本数据进行预处理,从而实现知识的数学表示。在知识聚合过程中应充分挖掘隐藏在文档中的知识,寻找知识之间特有的关联。谱聚类的方法可以将文本内容进行划分,并且发现文本之间的关系以及文档内容中包含的知识。本文将提取的文本特征词作为概念,计算概念的相似度,用改进的语义相似度矩阵代替空间向量模型,通过概念的语义相似度构建文本相似度矩阵,将其作为谱聚类的输入矩阵,利用谱聚类作为知识聚合的方法,从而降低矩阵的高维度,提高聚类结果的准确性。本文构建的基于谱聚类的虚拟健康社区知识聚合方法模型如图 1 所示:

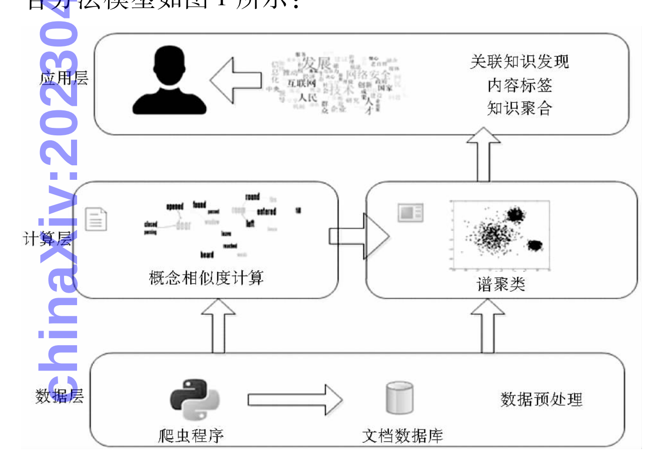


图 1 虚拟健康社区知识聚合方法模型

如图 1 所示,虚拟健康社区知识聚合方法模型由数据层、计算层以及应用层 3 个层次构成。数据层通过爬虫程序爬取虚拟健康社区中的主题帖子,将其以文本的形式保存在数据库中。利用分词软件对内容进行分词,并统计词频。通过筛选得到能够表明该帖子内容的特征关键词即概念。从而完成知识的数学形式表示。在计算层中,计算各个概念的相似度,进而得到知识主题的相似度。将其作为相似度矩阵,对其进行计算,得到拉普拉斯矩阵,进而对其进行谱聚类。通过谱聚类,可以有效发现知识资源间存在的语义关联。在应用层中,利用基于知识主题相似度的谱聚类方法对文本中的知识进行聚合,从而发现各个文档之中的相似性以及其中蕴含的知识,从而有效提高健康虚拟社区的知识服务质量以及用户满意度。

3.1 概念相似度计算

概念是指从文档中提取具有专指性且能反映文档主题的词语或短语,该词语既能体现文档的核心又可以体现文档的主题知识,还可以覆盖文档的内容,方便用户对文档进行索引和查找。通过概念提取,使得用户能够更加清晰直接地了解文本知识的内容和总体概貌,因此,本文将提取的关键词作为虚拟健康社区帖子中蕴含的知识概念进行表示和计算。概念是对虚拟健康社区帖子知识的表示,通过对概念相似度的计算,可以获得虚拟健康社区帖子知识之间的关系。在虚拟健康社区中,用户交流概念的共现关系体现了知识的潜在关系。两个帖子中共同出现的概念越多,说明这两个帖子的内容越相近。这个模型较为简单,但可以在一定程度上满足应用的需求,至今在文本信息检索、文本数据挖掘等领域被广泛应用[18]。因此,共现关系可以用于计算概念相似性并作为知识聚合的基础。

目前计算相似度的方法包括基于内容的相似度计算方法、基于属性的相似度计算方法、基于距离的相似度计算方法。由于属性是事物本身的内在特征,用户通常使用属性对事物进行辨识。这样,利用属性相似度计算公式可以有效对事物进行区分,并且基于事物本身的属性可以体现事物之间的关联程度。通常使用如下公式计算属性相似度:

$$Sim(K1,K2)=\frac{f(K1\cap K2)}{[f(K1\cap K2)+\alpha * f(K1-K2)+\beta * f(K2-K1)]}$$

公式(1)

在虚拟健康社区知识聚合方法中,使用虚拟健康社区帖子之间的概念来对其进行衡量。在公式(1)中 $K1\cap K2$ 表示概念在帖子中共同出现的次数,$K1-K2$ 表示 K1 出现而 K2 不出现的次数,$K2-K1$ 表示 K2 出现而 K1 不出现的次数。为了简便计算,将 α、β 系数都设定为 0.5。

3.2 知识主题相似度计算

知识主题相似度是指文本间主题或内容的相似程度,一般通过提取文本的特征关键词或概念进行计算[19]。与文本相似度概念类似,在计算两个帖子之间的知识主题相似度时,可以通过提取帖子中的概念或特征关键词,将帖子表示成概念的集合形式,即词向量的形式[20],进而通过包含的概念对其进行描述,方便计算相似度。

在计算知识主题相似度时,需要计算知识主题之间的语义距离。计算帖子之间文本距离的公式如下所示:

$$\mathrm{Dist}(d_x,d_y)=\mathrm{Dist}(\Lambda_{i=1}^{n}K_{xi},\Lambda_{j=1}^{m}K_{yj})=\frac{1}{d}\sum_{i=1}^{n}\sum_{j=1}^{m}f_i\times f_j\times\frac{1}{1+\mathrm{Sim}(K_i,K_j)}\quad 公式(2)$$

其中 d_x、d_y 为两个不同帖子,xi、yj 分别为帖子 d_x、d_y 所包含的概念;f_i、f_j 分别表示概念 xi、yj 在帖子 d_x、d_y 中出现的次数。n、m 分别为两个帖子所包含概念的个数。d 为在两个帖子中出现的概念的数量的和。这里使用 d 的目的是考虑到某一概念在帖子中出现的次数过多,导致帖子语义距离过大,因此使用 d 对该距离进行归一化。

综上,本文将知识主题语义相似度定义如下:

$$\mathrm{sim}(d_x,d_y)=\frac{1}{1+\mathrm{Dist}(d_x,d_y)}\quad 公式(3)$$

从公式(3)中可以看出,语义距离越大,知识主题间的相似度越小。

3.3 基于谱聚类的知识聚合算法

本文基于相似度矩阵的谱聚类算法提出了虚拟健康社区知识聚合方法。该方法抽取虚拟健康社区帖子的关键词,作为虚拟健康社区的概念。用概念列表表示知识主题,两个知识主题之间的相似度就可以转化为求解概念之间的相似度。通过计算概念之间的相似度,得到虚拟健康社区两个帖子之间的相似度。将其作为相似度矩阵,通过计算得到拉普拉斯矩阵,进而对其进行谱聚类。通过谱聚类,可以有效发现虚拟健康社区帖子间存在的语义关联。

虚拟健康社区知识聚合算法的描述:

输入:n 个数据点,聚类的个数 K

输出:聚类结果 C (C1,C2,…,Cn)

方法:

Begin

①构造相似性矩阵 $W\in R^{n\times n}$;构造度矩阵 $D\in R^{n\times n}$;

②改造拉普拉斯矩阵 L = D-W;

③求出 L 前 k 个特征值及其所对应的特征向量,将 k 按照特征值大小进行排序;并构造特征向量 V;

④将 V 看做是 k 维空间的一个向量,其中 $y_{ij}=v_{ij}/(\sum_j v_{ij}^2)^{1/2}$,使用聚类方法进行聚类。

End

谱聚类是一种基于图划分的聚类方法。通常是将数据放入到无向图中,利用数据点之间的权重,求得图的邻接矩阵。通常距离较远的点之间权重值较低,而距离较近的点之间权重较高。这样可以将权重作为相似度来衡量点之间的相似性。这里包括全连接、近邻连接等方法。本文计算虚拟健康社区帖子之间的相似度,由此构造相似度矩阵。这个相似度矩阵是一个对称矩阵。这样就进一步得到了度矩阵。通过计算,可以得到拉普拉斯矩阵 L。求得 L 的 k 个特征值,并且按照 k 的大小构造特征向量 V。把 V 的每一行都看作是新的数据,这样就可以使用聚类方法进行划分,从而得到聚类结果 C (C1,C2,…,Cn)。谱聚类只要求数据之间的相似度矩阵,这种处理方法实际对矩阵进行了降维,有利于处理稀疏的数据。

4 实验过程及结果

好大夫在线是一家深受患者信赖的互联网医疗平台。它在保证提供标准化高质量医疗服务的基础上将互联网思维及技术融入其中,探索出一套"网络咨询与线上答疑相同步、在线转诊与复诊相结合、预约专家门诊与签约私人医生相配套"的医疗服务模式,既方便了医患之间的沟通,又有效地扩大了自身影响力与权威性,为缓解当前紧张的医患矛盾提供了一个新方向。用户通过在平台上的经验交流与分享,逐步形成了极具影响力的医疗学术论坛,为进一步提高医疗服务的质量、加强线上线下服务结合的紧密度打下坚实基础。因此,本文利用好大夫在线的数据进行算法验证,通过 Python 编程爬取了心血管内科常见疾病标签下的文章内容共计 800 篇。数据预处理是对数据进行简化的过程,对数据进行去除停用词、分词、减噪等处理,提取出实验需要的、满足一定格式的数据内容,利用 Python 的 Jieba 功能进行分词和词频统计,计算相似度值并最终形成聚类。

4.1 概念抽取及相似度计算

通过自然语言处理可以从文本数据中提取知识,由于这些知识往往具有特定的结构和模式,可以将这些知识作为概念进行计算[21－22]。在知识聚合过程中,概念为知识聚合提供了最细粒度知识单元[23]。健康虚拟社区中,用户进行交流的内容往往围绕着某个领域的特定问题进行展开,这时从内容中抽取的概念往往可以代表该领域的知识,再对概念进行聚合就可以获得相似的知识。在进行知识聚合之前,需要对获取的概念进行整理,如利用领域相关度和一致度计算公式来剔除一些无关的概念或者无意义的概念[24]。领域相关度计算公式如下:

$$DR(t_i, D) = P(t_i \mid D) / \sum_{i=1}^{n} P(t_i \mid D)$$

公式(4)

其中 $P(t_i \mid D) = freq / \sum_{i=1}^{n} freq_i$，freq 为候选概念出现的频率。领域一致度的计算公式如下：

$$DC(t_i, D) = \sum_{i=1}^{n} P(t_i \mid D) \times \log \frac{1}{P(t_i \mid D)}$$

公式(5)

可得概念的抽取公式如下：

$$TW_i = \alpha \times DR(t_i, D) + \beta \times DC(t_i, D)$$ 公式(6)

为计算简便，将 α、β 设置为 0.5，据此可以对获取的概念进行整理并得到相关概念。

好大夫在线包括疾病介绍、病因症状、预防检查、疾病诊疗、就诊指南、护理保健等栏目。本文选取了所有标签下的分组内容进行相似度计算，利用 Python 语言编程进行预处理与编程计算，计算结果如图 2 所示：

```
[[0.36997203504475618, 0.3900822208524119, 0.40414117241292369, 0.43411424010295679, 0.47024538370713831, 0.51921339927454846, 0.5429
389397910196908, 0.97463105170256614, 1.3434774227511539, 2.4756972643352895], [0.38887041508987474, 0.39889086087515341, 0.425461051
7405934662, 0.60440861806887713, 0.6967064760981142, 0.73046830478388958, 0.91456999942610273, 1.1201878824744236, 2.1616590593665617
171, 0.99600386789982265, 1.0111439584875326, 1.0914172529321418, 1.1250268590925556, 1.1925714114967294, 1.2591962539210548, 1.38800
53110306216, 0.51184308291198644, 0.52462776744333128, 0.56565033020867639, 0.59712384657303852, 0.61640227436434525, 0.655114209289
1306, 1.5937243856736936, 1.8587340221894866, 2.2665437783630615], [0.4037052247149101, 0.40487462376220756, 0.42880960200124685, 0.4
52368039670470601, 0.53806405634412446, 0.62862404062291488, 0.72434763177784578, 0.77930476739667043, 1.060929203276516l, 5.05630596
07148587408, 0.86820310489354546, 0.99439780374481257, 1.0309513921590168, 1.1221433400598078, 1.2301473359719899, 1.370003861298543,
.84633324637186524, 0.84672621069731946, 0.85687901431824298, 0.90761094361240935, 0.90779056949179393, 0.90804373794953952, 0.924548
97988636, 1.2415083509512814, 1.578620349963513], [0.81268229105945333, 0.84810968860127822, 0.96474069414558095, 1.0169767601590245,
36865221924441O6, 1.4709339078152668, 1.5682178659448922, 1.6543025765997803, 1.7600038019972388, 3.536000456991689]]
d xin zhongxin [[0.36997228938968063, 0.3900830534774436, 0.40414221351450208, 0.4341148598587718, 0.47024548949580658, 0.51921339098
15260166, 0.86389615501748518, 0.97463089284778726, 1.3434902179502008, 2.4757179666080238], [0.38887170471326898, 0.39889232113861322
636, 0.57976196692194315, 0.60440645826169259, 0.69670402751198135, 0.73046501151690002, 0.91456824096337208, 1.1201806720080885, 2.1
.9680938312816747, 0.99531986042069776, 1.0104398487456705, 1.0906755030480473, 1.124481641616806, 1.1920354397051336, 1.258667017319
3], [0.50638533850312772, 0.51185079226868924, 0.52463384254179113, 0.56566017019136894, 0.59713019085969143, 0.61641045445091491, 0.
.2310635042521236, 1.5937832413246544, 1.8588061131530937, 2.2666076570767952], [0.40170553076866766, 0.40240347967104306, 0.42624447
88906584409, 0.51895732929242455, 0.53273419134074784, 0.62295153330739605, 0.71794278751173701, 0.77265104741905111, 1.0456949685540
12047, 0.84382360450299O1, 0.86821110306654303, 0.99435215532058796, 1.0309017485269001, 1.1220954241390031, 1.2300756716005625, 1.36
094266478652, 0.84664792672009526, 0.84697237307116546, 0.85703389898052085, 0.90772691596075106, 0.907798766312505, 0.90790003369560
31, 1.1470268856506611, 1.2421954896211767, 1.579719442284127], [0.81288611616435469, 0.8483070287563852, 0.96494811191553009, 1.0171
900026253, 1.3690191677983476, 1.4713597526476716, 1.5686726061298528, 1.6547845941101318, 1.7605455885823411, 3.5367114784334759]]
```

图 2 部分概念相似度计算结果

4.2 知识主题聚类实验及可视化结果

如何创建相似度矩阵，使其更加真实地反映数据点之间的近似关系，使得相近点之间的相似度更高，相异点之间的相似度更低，是谱聚类算法必须要解决的一个问题。高斯相似函数是经典谱聚类算法中计算两点间相似度的常用方法。使用高斯核函数时，相似矩阵和邻接矩阵相同。在 Python 中实现谱聚类的算法时，也可以选择高斯核函数进行。一般需要对高斯核函数中的参数 n_clusters 和 gamma 进行调参，选择合适的参数值。在本方法中，分别考虑当聚类数量 n_clusters = 3、4、5、6 这 4 种情况，对应的 gamma 选择 0.01、0.1、1、10 等 4 种情况，具体计算得分值如图 3 所示：

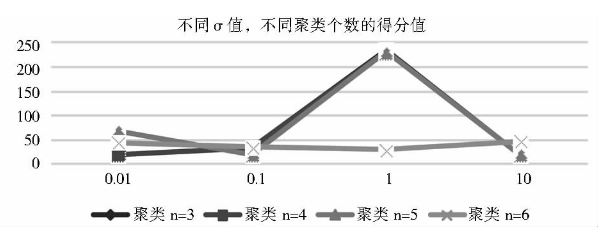


图 3 不同 σ 值，不同聚类个数的得分示意

对于不同的聚类结果，最高得分为 234.67，此时，n_clusters 是 5，而 gamma 是 1 或者 0.1。

本文的知识聚合结果如图 4 所示：

```
result - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
['A46.txt', 'A47.txt', 'B99.txt', 'E12.txt', 'E2.txt', 'E26.txt', 'E28.txt', 'E29.txt', 'E67.txt', 'E78.txt', 'E80.txt'
'C97.txt', 'E94.txt'
'B98.txt', 'E18.txt', 'E19.txt', 'E25.txt', 'E65.txt', 'E68.txt', 'E84.txt', 'F44.txt', 'K22.txt', 'K24.txt'
'C96.txt', 'D9.txt', 'E16.txt', 'E30.txt', 'H37.txt'
'C38.txt', 'E21.txt', 'E64.txt', 'E77.txt', 'K34.txt'
'C39.txt', 'E73.txt', 'E74.txt', 'F45.txt'
'A48.txt', 'E11.txt', 'E20.txt', 'E31.txt', 'E32.txt', 'E54.txt', 'E6.txt', 'E63.txt', 'E66.txt', 'E75.txt', 'E76.txt', 'E86.txt'
'E0.txt', 'E13.txt', 'E55.txt', 'E62.txt', 'E69.txt', 'E71.txt', 'E72.txt', 'H36.txt', 'K33.txt'
'B49.txt', 'B73.txt', 'D1.txt', 'D10.txt', 'D12.txt', 'D31.txt', 'D37.txt', 'D38.txt', 'D52.txt', 'D81.txt', 'D89.txt', 'D90.txt', 'D99.txt',
'F36.txt', 'F43.txt', 'F45.txt'
'C4.txt', 'C80.txt', 'D100.txt', 'D11.txt', 'D16.txt', 'D17.txt', 'D2.txt', 'D20.txt', 'D24.txt', 'D32.txt', 'D35.txt', 'D50.txt', 'D57.txt',
'D6.txt', 'D60.txt', 'D64.txt', 'D70.txt', 'D71.txt', 'D88.txt', 'D92.txt', 'D94.txt', 'E54.txt', 'J48.txt'
'D26.txt'
'C27.txt', 'C65.txt', 'D13.txt', 'D25.txt', 'D41.txt', 'D62.txt', 'D67.txt', 'E23.txt', 'E55.txt', 'E83.txt', 'F42.txt', 'H84.txt'
'D63.txt', 'D8.txt', 'D95.txt', 'E77.txt', 'F19.txt', 'F39.txt', 'H14.txt'
'D87.txt', 'F46.txt']
['C33.txt', 'D75.txt', 'D97.txt', 'E22.txt', 'E56.txt', 'F61.txt', 'F79.txt', 'H59.txt', 'H74.txt'
'D9.txt', 'E78.txt', 'E82.txt'
'D28.txt', 'D29.txt', 'D34.txt', 'D40.txt', 'D47.txt', 'D53.txt', 'D98.txt', 'F5.txt', 'J30.txt', 'J86.txt'
'D21.txt', 'D58.txt', 'D72.txt', 'D85.txt', 'D91.txt', 'D93.txt', 'F44.txt', 'H66.txt', 'J76.txt'
'A10.txt', 'B2.txt', 'B20.txt', 'B4.txt', 'B40.txt', 'B41.txt', 'B44.txt', 'B45.txt', 'B72.txt', 'C52.txt', 'C56.txt', 'C66.txt'
'A12.txt', 'A24.txt', 'A27.txt', 'A36.txt', 'A37.txt', 'A8.txt', 'A9.txt', 'B18.txt', 'B19.txt', 'B3.txt', 'B30.txt', 'B34.txt', 'B71.txt', 'C59.txt'
'A11.txt', 'A35.txt', 'A69.txt', 'B0.txt', 'B1.txt', 'B13.txt', 'B6.txt', 'B7.txt', 'C58.txt'
'B21.txt', 'B23.txt'
'B22.txt', 'B29.txt', 'B33.txt', 'B62.txt', 'C49.txt', 'C63.txt', 'C64.txt'
'B28.txt', 'B31.txt', 'B32.txt', 'B43.txt', 'B46.txt', 'B47.txt'
'A26.txt', 'A38.txt', 'A70.txt', 'B15.txt', 'B39.txt', 'C50.txt', 'C51.txt', 'C53.txt', 'C54.txt', 'C57.txt', 'C61.txt'
'A25.txt', 'A48.txt', 'A68.txt', 'B14.txt', 'B16.txt', 'B55.txt', 'C60.txt'
'C18.txt', 'H72.txt', 'H92.txt', 'K10.txt', 'K12.txt', 'K17.txt', 'K19.txt', 'K47.txt', 'K7.txt', 'K79.txt'
'H45.txt', 'K14.txt', 'K26.txt', 'K40.txt', 'K53.txt', 'K71.txt', 'K95.txt', 'K97.txt'
'C27.txt', 'H2.txt', 'H23.txt', 'H28.txt', 'K100.txt', 'K46.txt', 'K85.txt', 'K98.txt']
```

图 4 部分知识聚合结果示意

如图 4 所示，本文将文档进行标号。为了更好地对文档进行区分，采用"字母 + 数字"的格式。文档完成聚类划分后，采用中括号的形式进行区分。在分词时，往往需要根据用户所在的领域进行分词，并且分词时包含领域术语。Jieba 划分时，往往存在过度划分的情况，可以利用停用词提高信息检索时的搜索效率以及节约存储空间。停用词一般是由人工输入的，针对用户所在领域的专用术语，最终构造一个停用词表。本文利用 Python 中的库函数 jieba. load_userdict(file_

name）加载停用词表。

获取概念是知识聚合的基础，获取概念后针对概念进行处理，以抽取的概念为对象，以基于属性关系的相似度作为知识聚合的计算依据，实现虚拟健康社区的知识聚合。聚合后为了更好地展示聚合效果，本文将虚拟健康社区知识聚合的结果通过词云进行显示，见图 5。据图 5 中有关“心血管疾病”的相关知识聚合结果可以看出，将其分为 5 类较为合理。相关的知识主题包括瓣膜病变、高血压、心律失常、先天性心脏病、心绞痛，进一步可以通过知识聚合结果发现相关知识。例如高血压知识主题标签中，高血压与血管壁压力有关；从心律失常知识主题标签中可以看到心律失常有关的词语包括频率、节律、起源、异常等词语，心律失常也包括窦性、逸搏、异位等产生现象；冠心病心绞痛是指由于冠状动脉粥样硬化狭窄导致冠状动脉供血不足，等等。

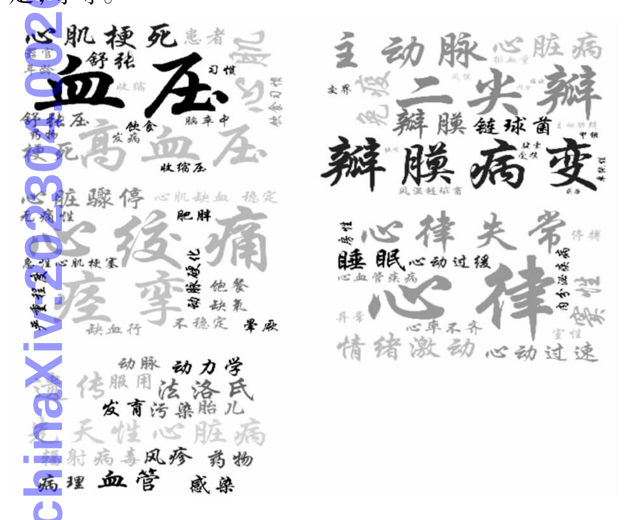


图 5　虚拟健康社区知识聚合结果词云

在好大夫在线网站中，也采用了主体导航的方式对疾病进行分类。但是这些分类较为分散，不能有效聚焦主题。用户在浏览相关疾病时，耗费大量精力寻找与自己需求相关的疾病，导致用户迷失在海量的知识中。本文的知识聚合结果是按照知识主题进行分类，通过概念的词频进行展示，可以作为选择相关主题的依据。例如冠心病心绞痛可能出现的症状是心脏骤停、痉挛或者晕厥，产生的原因可能是肥胖、心肌缺血等。另外饱餐和缺氧也可能引起此类疾病。用户可以根据出现的概念选择所要了解的主题。另外，在该聚合结果中也可以看到，与某一领域相关的概念已经被完全展示出来，用户可以根据需求获得更多的选择。如果用户已经表明对某类话题感兴趣，那么可以通过聚合结果向用户推荐更多主题。例如与心律有关的其他领域概念可以为心律失常或者心律不齐等。如果用户关注了心律这个领域概念，那么可以根据聚合的结果，为用户推荐相关的领域概念或者话题，从而更加有针对性地为用户提供服务。

通过以上研究可以发现基于谱聚类的虚拟健康社区知识聚合方法具有一定的可行性和有效性，可以帮助用户了解该话题中的相关知识以及相关主题，用户通过主题簇可以迅速查找相关知识内容。通过知识聚合方法，可以帮助虚拟健康社区进一步完善知识检索、知识发现、知识导航等服务，也可以基于该方法实现知识推荐、知识图谱等功能。

5　结语

本文提出了基于谱聚类的虚拟健康社区知识聚合方法。首先，爬取好大夫在线虚拟健康社区网站的内容，通过分词软件 Jieba 对文本进行预处理，提取出可以代表虚拟健康社区知识的概念。利用相似度计算公式计算概念相似度，以此为基础构造虚拟健康社区帖子知识主题的相似度矩阵。对相似度进行规范化，并通过计算得到拉普拉斯矩阵。求得前 k 个特征值，并且按照 k 的大小构造特征向量 V，把 V 的每一行都看做是新的数据，这样就可以使用聚类方法进行划分，从而得到知识聚合结果。利用谱聚类的方法对虚拟健康社区中的知识进行聚合，可以帮助虚拟健康社区用户迅速了解相关知识主题及知识内容，并且可以为虚拟健康社区用户提供具有针对性的知识服务，从而帮助虚拟健康社区有效提升用户体验和服务质量。

参考文献：

[1] 新华社. 我国社会主要矛盾转化的背后[EB/OL]. [2019－06－26]. http://cpc.people.com.cn/19th/n1/2017/1021/c414305-29600806.html.

[2] HAJLI M N. Developing online health communities through digital media[J]. International journal of information management, 2014, 34(2):311－314.

[3] 毕强. 数字资源：从整合到聚合的转变[J]. 数字图书馆论坛, 2014(6):1.

[4] 李亚婷. 知识聚合研究述评[J]. 图书情报工作, 2016, 60(21):128－136.

[5] 王敬东. 基于知识聚合的数字图书馆信息智能检索模型[J]. 图书馆学研究, 2014(21):72－76, 71.

[6] 贯君，毕强，赵夷平. 基于关联数据的知识聚合与发现研究进展[J]. 情报资料工作, 2015(3):15－21.

[7] 李洁. 基于 SNA 的馆藏数字资源知识聚合可视化研究[D]. 长春：吉林大学, 2016.

[8] 陈果，朱茜凌，肖璐. 面向网络社区的知识聚合：发展、研究基础

与展望[J]. 情报杂志,2017,36(12):193－197,192.
[9] 张连峰,李慧,逄云鹤. 基于虚拟学术社区的知识聚合模型构建研究[J]. 情报科学,2019,37(6):55－60,74.
[10] 胡媛,刁首琪,朱益平,等. 基于知识聚合的数字图书馆社区服务推送系统设计与实现[J]. 情报科学,2017,35(11):72－77.
[11] 商宪丽,王学东,张煜轩. 基于标签共现的学术博客知识资源聚合研究[J]. 情报科学,2016,34(5):125－129.
[12] LIANG K, WANG C, ZHAN Y Y. Knowledge aggregation and intelligent guidance for fragmented learning[J]. Procedia computer science, 2018,131(4): 656－664.
[13] TARKO V, DRAGOSALIGICA P. From "Broad Studies" to Internet-based "Expert Knowledge Aggregation". Notes on the methodology and technology of knowledge integration[J]. Futures, 2011, 43(9):986－995.
[14] RITOU M, BELKADI F, YAHOUNI Z, et al. Knowledge-based multi-level aggregation for decision aid in the machining industry[J]. CIRP annals, 2019,68(1):475－478.
[15] OOSTERMANA J, YANG J, ALESSANDRO B, et al. On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks[J]. Computer networks, 2015, 90(29): 133－149.
[16] JANANI J, VIJAYARANI S. Text document clustering using spectral clustering algorithm with particle swarm optimization[J]. Expert systems with applications, 2019,134(15): 192－200.
[17] LI X, WANG Z J, HU R L, et al. Recommendation algorithm based on improved spectral clustering and transfer learning[J]. Pattern analysis and applications, 2019, 22(2):633－647.
[18] 李航. 统计学习方法[M]. 2 版. 北京:清华大学出版社,2019.
[19] 王春柳,杨永辉,邓霏,等. 文本相似度计算方法研究综述[J]. 情报科学,2019,37(3):158－167.
[20] 李枫林,柯佳. 基于深度学习的文本表示方法[J]. 情报科学,2019,37(1):156－164.
[21] LIU K H, HOGAN W R, REBECCA S. Crowley. Natural language processing methods and systems for biomedical ontology learning [J]. Journal of biomedical informatics 2011, 44(1):163－179.
[22] ANDRÉSPAREDES-VALVERDE M, ÁNGELRODRÍGUEZ-GARCÍA M, RUIZ-MARTÍNEZ A, et al. ONLI: an ontology-based system for querying DBpedia using natural language paradigm[J]. Expert systems with applications, 2015, 42(12): 5163－5176.
[23] FRANTZI K T, ANANINADOU S. The C-Value/NC-Value domain independent method for multi-word term extraction[J]. Journal of natural language processing, 2008,6(3):145－179.
[24] 廖福燕. 本体构建中概念和关系获取方法研究[D]. 西安:西安建筑科技大学,2011.

作者贡献说明:
张海涛:研究思路与方法拟定、数据分析、论文修订;
宋拓:数据采集、分析处理,论文初稿撰写;
周红磊:数据收集与整理;
张鑫蕊:论文修订。

Research on Knowledge Aggregation Method of Virtual Healthy Community Based on Spectral Clustering

Zhang Haitao[1,2] Song Tuo[1] Zhou Honglei[1] Zhang Xinrui[1]
[1] School of Management, Jilin University, Changchun 130022
[2] Jilin University Information Resource Research Center, Changchun 130022

Abstract: [**Purpose/significance**] To improve the quality of knowledge aggregation in healthy virtual communities and provide technical method support for virtual healthy community services. [**Method/process**] The method of spectral clustering was applied to knowledge in the virtual healthy community was extracted, and the semantic similarity matrix of the text was obtained by using the keyword co-occurrence. The spectral clustering was performed according to the text semantic similarity matrix, and the text was aggregated into text clusters. [**Result/conclusion**] The information published by the doctor's online health consultation platform was used as a data source for method validation. The results show that when the number of clusters is 5, the proposed method has the highest score. This method of spectral clustering considers the semantic relationship between words, fully exploits the potential information of virtual healthy community, improves the quality of knowledge aggregation, and provides a new way for knowledge aggregation and knowledge service.

Keywords: virtual healthy community knowledge aggregation spectral clustering similarity